Gappy Data: To Krig or Not To Krig?

Hasan Gunes [a], Sirod Sirisup [b], George Em Karniadakis [b,*]

[a] *Department of Mechanical Engineering, Istanbul Technical University, Gumussuyu, 80191 Istanbul, Turkey*

[b] *Division of Applied Mathematics, Brown University, Providence, RI 02912, USA*

Abstract

Data recovery and reconstruction methods for unsteady flow fields with spatio-temporal missing data are studied based on proper orthogonal decomposition (POD) and on Kriging interpolation. It is found that for sufficient temporal resolution, POD-based methods outperform Kriging interpolation. However, for insufficient temporal resolution, large spatial gappiness or for flow fields with black zones, Kriging interpolation is more effective. The comparison is performed based on randomly generated laminar and turbulent flow fields obtained from simulations of uniform flow past a circular cylinder.

Key words: Kriging, proper orthogonal decomposition, unsteady flow

1 Introduction

Data assimilation is routinely used in atmospheric and ocean modeling, however, this is not the case with more classical fluid mechanics problems in laboratory or open field applications. The recent rapid developments in quantitative imaging techniques, e.g., particle image velocimetry (PIV) and magnetic resonance imaging (MRI), and the simultaneous advances in large-scale simulation offer the possibility for integrating seamlessly flow simulations and experiments. A key element in this integration is the reconstruction of flow fields from a finite number of PIV or MRI images and the evaluation of error in such reconstruction. Even for relatively simple laboratory experiments such as the classical flow past a cylinder, information may be missing due to "shadowing" (i.e., obstructed view) and proximity to the cylinder walls or the frequency of the measurements may be below a certain threshold, e.g. in MRI

* Corresponding Author: G.E. Karniadakis, gk@cfm.brown.edu

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2005		**00-05-2005 to 00-05-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Gappy Data: To Krig or Not To Krig?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Brown University,Division of Applied Mathematics,182 George Street,Providence,RI,02912**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**30**	

measurements. Therefore, we have to work with gappy data where the spatio-temporal regions of missing data are known in advance or where missing data occurs at random.

The gappy data problem is not new as researchers have been working on it for many decades and many different statistical approaches are used in practice [1,2]. Local Kriging is an effective statistical method which has been used with success in geology and other fields to interpolate randomly scattered data. Unlike other estimation procedures, Kriging provides a measure of the error and associated confidence in the estimates [3]. It is also appropriate for the so-called "black zones", i.e. regions of the domain that may miss data at *all* times. A non-statistical method for spatio-temporal gappiness, based on proper orthogonal decomposition (POD), was first proposed in [4]; it was extended and tested in flow problems in [5,6]. Independently, another version of the method was presented in [7] for oceanographic data using again the method of empirical orthogonal functions (EOF) – a method conceptually identical to POD. In particular, the versions presented in [6,7] make this non-statistical approach more robust compared to its original formulation in [4] as the extended method does not depend on the initial guess of the flow field in the gappy regions.

In the current work we present a comparative study between the aforementioned two POD-based reconstruction procedures against the Kriging reconstruction procedure. Of particular interest is the *maximum* possible resolution that can be obtained given a certain gappiness level. This, in general, is a complex question and the answer depends critically on the type of spatio-temporal gappiness in the flow field, the smoothness of the data, and the Reynolds number. To this end, we have selected the flow past a cylinder as a test problem due to the previous experience with this flow that shows that a low-dimensional representation indeed exists, see [6,8–11], and thus redundancy in the available gappy data can be exploited for an accurate reconstruction. Here we consider both two- and three-dimensional gappy flow fields in the laminar and turbulent regimes and vary appropriately the available number of flow snapshots.

The paper is organized as follows. In the next section we present an overview of two versions of the POD based reconstruction that we will employ as well as of the Kriging interpolation. We then describe in section 3 the procedure we follow in producing gappy data sets based on direct numerical simulations (DNS). In section 4 we present the results, separately for two-dimensional laminar flows and for three-dimensional laminar and turbulent flows. We conclude in section 5 with a brief summary.

2 Methodology

We present first two versions of the POD-based reconstruction approach and subsequently a brief overview of the Kriging interpolation.

2.1 POD-based Reconstruction

The problem of reconstructing a gappy field using POD orthogonal modes was first considered by Everson & Sirovich in [4] for an image reconstruction static problem. The gappy field can be written in the following way:

$$\mathbf{u}_g(\mathbf{x},t) = \left(u(\mathbf{x},t)m_1(\mathbf{x},t), v(\mathbf{x},t)m_2(\mathbf{x},t), w(\mathbf{x},t)m_3(\mathbf{x},t)\right), \tag{1}$$

where $m_j(\mathbf{x},t)$ tracks the spatio-temporal missing data. It is defined by

$$m_j(\mathbf{x},t) = \begin{cases} 1 \quad \text{if the } j^{th} \text{ component of the field is known in } (\mathbf{x},t) \\ 0 \quad \text{if the } j^{th} \text{ component of the field is missing in } (\mathbf{x},t). \end{cases} \tag{2}$$

The procedure proposed by Everson & Sirovich completes the missing spatio-temporal dynamics starting from a certain initial guess for the unknowns and proceeds iteratively. At the heart of the method is the minimization of a functional defined in the spatio-temporal domain where the field is known. Let us denote by $\widetilde{\mathbf{u}}\left(\mathbf{x},t\right)$ a *completed* field based on some initial guess; the standard Everson-Sirovich method employs the *time average* value at that location as initial guess. Subsequently, we perform POD of $\widetilde{\mathbf{u}}\left(\mathbf{x},t\right)$ to obtain the *guessed* spatial and temporal modes. This decomposition has the form

$$\widetilde{\mathbf{u}}\left(\mathbf{x},t\right) = \sum_{k=1}^{N} \widetilde{\mathbf{\Phi}}_k\left(\mathbf{x}\right) \widetilde{\zeta}_k\left(t\right), \tag{3}$$

where $\widetilde{\zeta}_k(t)$ is the k^{th} guessed temporal mode and $\widetilde{\mathbf{\Phi}}_k(\mathbf{x})$ is the k^{th} guessed spatial mode. The proper functional for minimization is

$$\begin{aligned} F_g\left[\widetilde{\xi}_k\right] &= \left\| \widetilde{\mathbf{u}}\left(\mathbf{x},t\right) - \sum_{k=1}^{M} \widetilde{\mathbf{\Phi}}_k\left(\mathbf{x}\right) \widetilde{\xi}_k\left(t\right) \right\|^2_{\text{Gappy}} \\ &= \left(\widetilde{\mathbf{u}}\left(\mathbf{x},t\right) - \sum_{k=1}^{M} \widetilde{\mathbf{\Phi}}_k\left(\mathbf{x}\right) \widetilde{\xi}_k\left(t\right) \;,\; \widetilde{\mathbf{u}}\left(\mathbf{x},t\right) - \sum_{k=1}^{M} \widetilde{\mathbf{\Phi}}_k\left(\mathbf{x}\right) \widetilde{\xi}_k\left(t\right) \right)_{\text{Gappy}}, \end{aligned} \tag{4}$$

where the "Gappy" norm is defined on the support of $\widetilde{\mathbf{u}}(\mathbf{x},t)$, i.e. the spatio-temporal domain on which the values of $\widetilde{\mathbf{u}}(\mathbf{x},t)$ are known *for sure*. Also, M

is the number of modes that we use in the reconstruction process, which is different than the number of available snapshots N ($M \leq N$). Minimization of this functional (4) leads to the linear system of algebraic equations

$$\sum_{j=1}^{M} \left(\widetilde{\mathbf{\Phi}}_i (\mathbf{x}) \ , \ \widetilde{\mathbf{\Phi}}_j (\mathbf{x}) \right)_{\Omega_g(t)} \widetilde{\xi}_j (t) = \left(\widetilde{\mathbf{u}} (\mathbf{x}, t) \ , \ \widetilde{\mathbf{\Phi}}_i (\mathbf{x}) \right)_{\Omega_g(t)} \quad i = 1, ..., M. \quad (5)$$

The unknowns are the new (non-normalized) temporal modes $\left\{ \widetilde{\xi}_k(t) \right\}$; also, $\Omega_g(t)$ is the gappy spatial domain at time t. Note that the $M \times M$ matrix

$$\left[\widetilde{\mathbf{K}} \right]_{ij} = \left(\widetilde{\mathbf{\Phi}}_i (\mathbf{x}) , \widetilde{\mathbf{\Phi}}_j (\mathbf{x}) \right)_{\Omega_g(t)} \quad (6)$$

has time-dependent coefficients.

We now summarize the above algorithm, which is the first version of the POD approach that we present:

Algorithm POD-1

(1) Use time-average values as initial guesses at the locations $m_j(\mathbf{x}, t) = 0$ to obtain N snapshots of an initial complete field $\widetilde{\mathbf{u}}(\mathbf{x}, t)$.
(2) Perform POD of $\widetilde{\mathbf{u}}(\mathbf{x}, t)$ to obtain N guessed spatial modes $\left\{ \widetilde{\mathbf{\Phi}}_i (\mathbf{x}) \right\}$.
(3) Select the number of modes M to be employed in the reconstruction.
(4) Construct the matrix $\left[\widetilde{\mathbf{K}} \right]_{ij} = \left(\widetilde{\mathbf{\Phi}}_i (\mathbf{x}) , \widetilde{\mathbf{\Phi}}_j (\mathbf{x}) \right)_{\Omega_g(t)}$ and the vector $[\widetilde{\mathbf{f}}]_i = \left(\widetilde{\mathbf{u}} (\mathbf{x}, t) , \widetilde{\mathbf{\Phi}}_i (\mathbf{x}) \right)_{\Omega_g(t)}$.
(5) Solve the $M \times M$ linear system: $\widetilde{\mathbf{K}} \widetilde{\xi} = \widetilde{\mathbf{f}}$ for the unknowns $\left\{ \widetilde{\xi}_k(t) \right\}$.
(6) Construct a new vector field as follows:

$$\widetilde{\mathbf{w}} (\mathbf{x}, t) = \sum_{k=1}^{M} \widetilde{\mathbf{\Phi}}_k (\mathbf{x}) \, \widetilde{\xi}_k (t) \quad (7)$$

and overwrite the previous guess, i.e., set

$$\widetilde{\mathbf{u}} (\mathbf{x}, t) = \widetilde{\mathbf{w}} (\mathbf{x}, t) \quad \text{only if} \quad m_j(\mathbf{x}, t) = 0.$$

(7) Upon convergence stop, otherwise go to (2).

A computationally expensive element in the above algorithm is the solution of the least-squares system in each iteration in steps 4 and 5. An alternative approach is to bypass these steps and approximate $\widetilde{\xi}_k \approx \widetilde{\zeta}_k$. For clarity, we present the entire modified procedure below:

Algorithm POD-2

(1) Use time-average values as initial guesses at the locations $m_j(\mathbf{x},t)=0$ to obtain N snapshots of an initial complete field $\widetilde{\mathbf{u}}(\mathbf{x},t)$.
(2) Perform POD of $\widetilde{\mathbf{u}}(\mathbf{x},t)$ to obtain N guessed spatial modes $\left\{\widetilde{\mathbf{\Phi}}_i(\mathbf{x})\right\}$ and the corresponding temporal modes $\widetilde{\zeta}_k(t)$.
(3) Select the number of modes M to be employed in the reconstruction.
(4) Construct a new vector field as follows:

$$\widetilde{\mathbf{w}}(\mathbf{x},t)=\sum_{k=1}^{M}\widetilde{\mathbf{\Phi}}_k(\mathbf{x})\,\widetilde{\zeta}_k(t) \tag{8}$$

and overwrite the previous guess, i.e., set

$$\widetilde{\mathbf{u}}(\mathbf{x},t)=\widetilde{\mathbf{w}}(\mathbf{x},t) \quad \text{only if} \quad m_j(\mathbf{x},t)=0.$$

(5) Upon convergence stop, otherwise go to (2).

This approach has been used in [7] for oceanographic data using the singular value decomposition formulation.

POD-1 breaks down when the matrix $\widetilde{\mathbf{K}}$ is singular. This includes, for example, the case in which a snapshot is missing (i.e., $\Omega_g(t^*)=0$). Also, both algorithms (POD-1 and POD-2) break down if there are any spatial subregions for which the data is missing at *all* the times. In general, the optimum number of modes M_o in the reconstruction depends on the initial guess. Specifically, by *optimum number* of modes we mean the number M_o for which the error is the smallest among all possible converged reconstructions. A more robust version was presented in [6] that does not depend on the initial guess and, in addition, enhances accuracy significantly. The main steps of this iterative procedure are:

Iterative Procedure

(1) Perform POD-1 but employ only $M=2$ modes in the reconstruction.
(2) Use the converged result from the previous step as a new initial guess and apply POD-1 but now employ $M=3$ modes in the reconstruction.
(3) Proceed similarly for the n^{th} iteration until the obtained eigenspectrum does not change anymore.

Although more costly, results presented in [6] suggest that the iterative POD-1 procedure leads to the maximum possible resolution of the true eigenspectrum and thus of possible accuracy in reconstructing the flow field. More importantly, the final solution will only depend on the degree of gappiness and not on the initial guesses in the gappy subregions. A similar iterative procedure for POD-2 can be used by starting from $M=1$.

We have found that the same iterative procedure is effective when we apply POD-2, see [12]. In the present paper, we will employ the *iterative* extensions of algorithms POD-1 and POD-2.

2.2 Kriging Interpolation

We adopt the Kriging interpolation method based on the *correlogram* rather than the usual variogram, (see [13] for more information), implemented in the Matlab toolbox DACE, [14]. A brief overview is as follows:

Let S be the sampling grid containing known m points, i.e., $S = [s_1...s_m]$ with $s_i \in \mathbb{R}^n$. A known value of a function at those locations is denoted as $y_i \in \mathbb{R}^p$, where we assume that we have p functions at the point, i, e.g., $p = 3$ for a three-dimensional velocity vector field. Let $\mathbf{Y}$ be the matrix containing all those responses, i.e., $\mathbf{Y} = [y_1...y_m]$, $y_i \in \mathbb{R}^p$.

Let us assume for simplicity that $p = 1$. Then, for the set S of our sampling grid, we define the $m \times l$ interpolation matrix $\mathbf{F} \in \mathbb{R}^{m \times l}$ as

$$\mathbf{F}_{ij} = f_j(s_i),$$

where $f_j : \mathbb{R}^n \to \mathbb{R}$ are polynomial functions. We also define the correlation matrix $\mathbf{R} \in \mathbb{R}^{m \times m}$ by $\mathbf{R}_{ij} = \mathbf{R}(\theta, s_i, s_j)$ where θ is the correlation parameter that defines implicitly the correlation length. Then, the Kriging interpolation at an unknown point x is

$$\hat{y} = \mathbf{f}(x)\boldsymbol{\mu}^* + \mathbf{r}(x)\boldsymbol{\gamma}^*, \tag{9}$$

where $\mathbf{f}(x) = [f_1(x), f_2(x), ..., f_{\frac{(n+1)(n+2)}{2}}(x)]$ are the functions in the regression model, and

$$\boldsymbol{\mu}^* = (\mathbf{F}^T\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{R}^{-1}\mathbf{Y},$$
$$\mathbf{r}(x) = [\mathbf{R}(\theta, s_1, x)...\mathbf{R}(\theta, s_m, x)]^T,$$

with

$$\boldsymbol{\gamma}^* = \mathbf{R}^{-1}(\mathbf{Y} - \mathbf{F}\boldsymbol{\mu}^*).$$

From this last formula we see that for each unknown point x, we just need to compute the interpolation vector $\mathbf{f}(x)$ and the correlation vector $\mathbf{r}(x)$. We first specify the correlation matrix $\mathbf{R}(\theta, s_i, s_j)$ and provide a range for θ within which an optimization procedure is performed based on the variance estimate. This will give the optimum value θ^* and based on that value then $\mathbf{R}(\theta^*, s_i, s_j)$, $\boldsymbol{\mu}^*$ and $\boldsymbol{\gamma}^*$ will also be known.

For multiple functions ($p > 1$), equation (9) is valid for each column of $\mathbf{Y}$, i.e. equation (9) holds with $\boldsymbol{\mu}^* \in \mathbb{R}^{l \times p}$ and $\boldsymbol{\gamma}^* \in \mathbb{R}^{m \times p}$.

The regression model used in all of our studies is based on second-order polynomials. Specifically, let x_j be the j^{th} component of $x \in \mathbb{R}^n$, then:

$$\begin{aligned} f_1(x) &= 1 \\ f_2(x) = x_1, ..., f_{n+1}(x) &= x_n \\ f_{n+2} = x_1^2, ..., f_{2n+1} &= x_1 x_n \\ f_{2n+2} = x_2^2, ..., f_{3n} &= x_2 x_n \\ \dots \quad \dots \quad f_l(x) &= x_n^2 \end{aligned}$$

where $l = \frac{(n+1)(n+2)}{2}$.

Also, for the correlation kernel $\mathbf{R}$, we employ a *Gaussian correlation model* since the flow field is continuously differentiable. More precisely, $\mathbf{R}$ has the form :

$$\mathbf{R}(\theta, w, x) = \prod_{j=1}^{n} R_j(\theta, w_j - x_j)$$

where

$$R_j(\theta, w_j - x_j) = \exp(-\theta_j (w_j - x_j)^2).$$

In the last sub-section of section 4 we will also use the exponential correlation model for comparison, defined by

$$R_j(\theta, w_j - x_j) = \exp(-\theta_j |w_j - x_j|).$$

For more information regarding the computational aspects of the Kriging procedure, see [15].

In order to apply the Kriging procedure to our gappy data we need to distinguish two cases. First, we consider the case that we have no black zones at all. In this case, we apply the Kriging procedure element-by-element, where element here refers to the spectral element of the computational mesh. Second, we consider gappy data with black zones. In this case, we embed the black zone in a larger region, roughly twice the size of the black zone, so that is surrounded with data points of known values. We then apply Kriging on the larger region.

3 Data Gathering

We first consider *two-dimensional* flow past a circular cylinder at Reynolds number $Re = 100$ and $Re = 500$. The computational domain is shown in figure 1. Uniform flow is imposed at the inflow boundary Γ_1 and also on Γ_3 and Γ_4 while on Γ_2 the zero Neumann condition on velocity is imposed and the

pressure is set to zero. On the cylinder surface Γ_5 the no-slip boundary condition is prescribed. Converged solutions were obtained using the spectral/hp element method [16]. The domain is discretized into 412 triangular elements while 7^{th}-order Jacobi polynomial basis are used to obtain resolution independent solutions. Forty ($N = 40$) and fifty ($N = 50$) snapshots of solutions are obtained from the DNS for Re=100 and Re=500, respectively, corresponding to one full vortex-shedding period T_s. (For Re=100 we obtained $T_s = 5.99$ while for Re=500 we obtained $T_s = 4.54$ in convective time units.) Points in the domain are then randomly dropped from the data set to produce gaps in the data with 10%, 25% and 50% missing data in each snapshot.

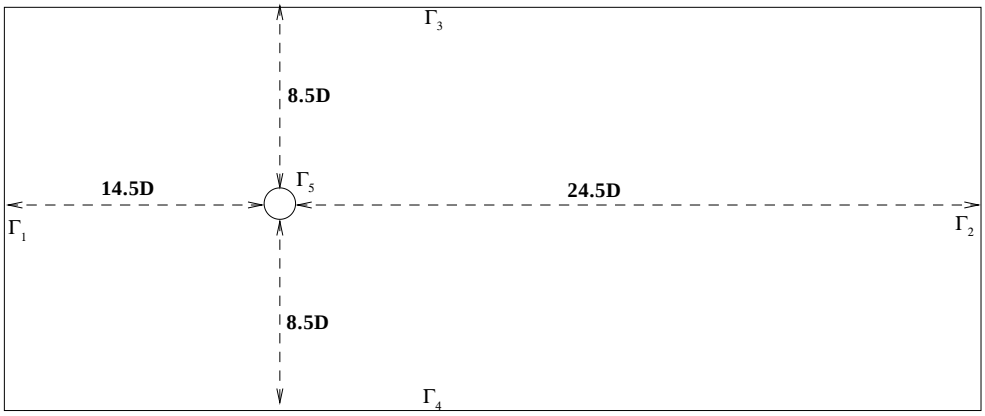

Fig. 1. Computational domain for flow past a circular cylinder.

Subsequently, we consider three-dimensional flow past a circular cylinder at Reynolds number, $Re = 185$ and $10,000$. The former corresponds to a periodic limit cycle (laminar flow with vortex shedding period $T_s = 5.5$ in convective time units) whereas the latter to a turbulent wake so we deal with "rough" data. A discretization similar to the two-dimensional cases is used here but with Fourier expansions employed along the spanwise location. In an effort to assess the spanwise convergence and its possible effect on the gappy data reconstruction procedures, we performed systematic refinement studies along the spanwise direction with $Z = 4, 8$ and $Z = 16$ Fourier collocation points along the span. In these simulations, the length of the cylinder span was kept constant at $\frac{L_Z}{D} = 4$, where D is the diameter of the cylinder, and the Reynolds number was kept constant at $Re = 185$ (based on the cylinder diameter). We refer to [17] for more information for the $Re = 185$ case and to [18] for the $Re = 10,000$; in the latter case a much higher resolution (about 15 million grid points) and systematic refinement studies were performed.

In order to generate the gappy flow field, we discard randomly the values of the solution on some nodes in every snapshot. In order to quantify the gappiness in a flow field, we define the "gappiness percentage" as the number of nodal data

points that are missing with respect to the total number of nodal data. Given a gappiness percentage, due to the distribution of the spectral elements, we actually discard much more data near the cylinder. Specifically, we consider flows fields with 25%, 50% and 75% gappy data. Due to random discarding process, these gappiness percentages vary slightly for each snapshot but are very close to the assigned values.

The gappy data set we employed were obtained from a fully three-dimensional spatio-temporal random distribution. In addition, in order to assess the effect of the spatial gappiness distribution, we have also considered "quasi three-dimensional" gappiness, that is, we keep the *same* gappiness distribution for all planes in the spanwise direction. Numerical results revealed that the quasi three-dimensional gappiness did not affect the reconstruction procedure in any significant way. Here we will not report such results separately.

4 Results

4.1 Two-Dimensional Gappy Fields

We will first present a comparison of results between the two versions of the POD-based reconstruction procedure (POD-1 and POD-2) for smooth and noisy gappy flow fields, and subsequently we will present results based on the Kriging interpolation. We separate these comparisons into two main categories corresponding to availability of data with high temporal resolution and with low temporal resolution. As we will demonstrate, temporal resolution is crucial in deciding which method is more effective in recovering complete fields.

4.1.1 Convergence

Given the similarity of the two versions of POD reconstruction we first compare their corresponding convergence rates for smooth and noisy data. For the cases we study here, since we know completely the original velocity field, we monitor the time-averaged relative error. In general, when we do not know the real field, we can adopt a cross-validation approach by taking out $1-3\%$ of *known* data and treating them like *missing* data. In figure 2 we plot the time-averaged error defined by

$$\sqrt{\sum_{i=1}^{M} \frac{1}{M} \frac{\int_\Omega (\mathbf{u}_i - \hat{\mathbf{u}}_i)^2 d\Omega}{\int_\Omega \mathbf{u}_i^2 d\Omega}}$$

versus the number of modes M employed in the reconstruction. This error decreases monotonically if the iterative procedure is applied, see figure 2 (right); however, without the iterative procedure presented in section 2 the error decreases initially but then diverges as the number of modes M employed in the reconstruction exceeds a certain value, see figure 2 (left). From the comparison of the results in figure 2, we see that the convergence of the POD-1 method is very close to the convergence of POD-2, however the latter is a much more efficient method. The results in the plot are for $Re = 500$ but similar results hold for $Re = 100$ not shown here. We note, however, that the non-monotonic decrease of the error is problem-dependent; for example, for some noisy data sets both POD-1 and POD-2 diverge above a certain number of modes employed in the reconstruction. This is demonstrated in figure 3, where we have superimposed Gaussian noise to the previous gappy data in the form:

$$\mathbf{Y} \mapsto \mathbf{Y} + \alpha \ \text{randn}(0,1),$$

where here we used $\alpha = 0.15$. We also note that in the presence of noise, POD-1, which involves solution of a least-square system, is more accurate than POD-2.

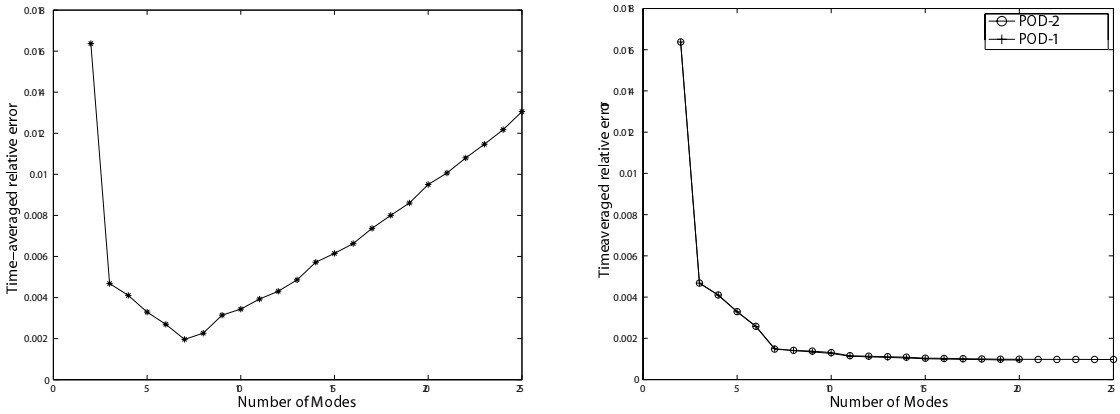

Fig. 2. Time-averaged error versus number of reconstruction modes of POD method for $Re = 500$ with 50% gappiness and no iterations (left) and that of POD-1 and POD-2 methods for $Re = 500$ with 50% gappiness (right).

4.1.2 High temporal resolution

Let us assume that we have available many snapshots of the flow field, e.g. $N = 50$ for $Re = 500$. We will compare the two POD versions against the results obtained with Kriging interpolation. In particular, since the domain is decomposed into non-uniform triangular elements with variable grid-spacing, we map each triangular element onto square elements with equal grid-spacing and then we apply the Kriging interpolation.

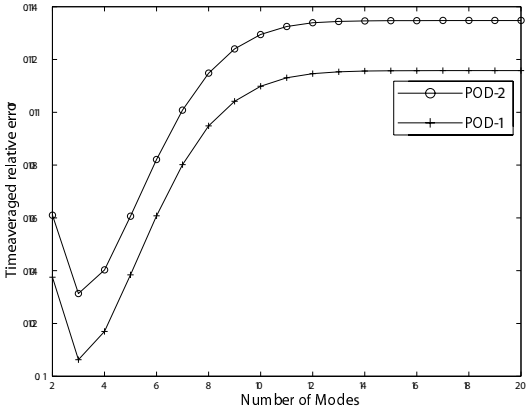

Fig. 3. Time-averaged error versus number of reconstruction modes of POD-1 and POD-2 methods for $Re = 500$ with 90% gappiness in the noisy data with $\alpha = 0.15$.

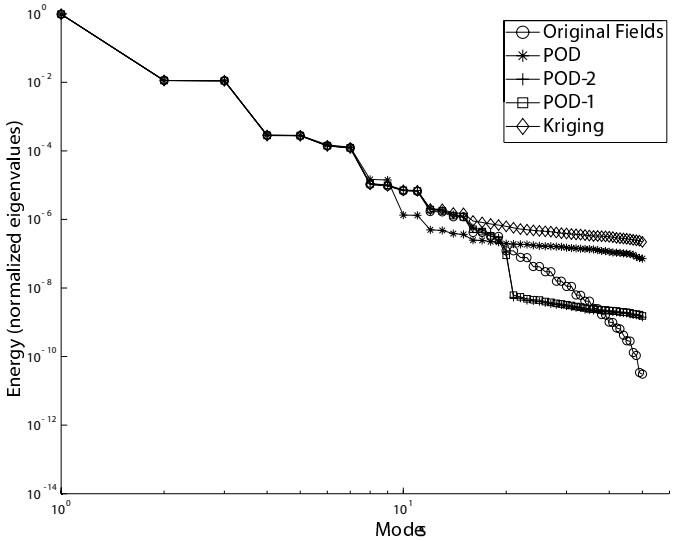

Fig. 4. Comparison of energy distribution of POD modes of the recovered data for $Re = 500$ with 50% gappiness.

We first compare the energy distribution corresponding to the different reconstructed fields, i.e., the normalized eigenvalues of the "reconstructed" POD modes. A typical result is shown in figure 4 for $Re = 500$ and 50% gappiness percentage. We see that the eigenspectra of POD-1 and POD-2 are in good agreement with the eigenspectrum of the complete original field up to mode $M_0 = 15$ whereas the Kriging interpolation produces an eigenspectrum accurate up to $M_0 = 10$. We have also included in the plot an additional result (denoted as "POD" in the legend) corresponding to POD-2 but without any iterations. It shows poor agreement with the complete data eigenspectrum, thus indicating the importance of applying the POD procedures *iteratively* up to full convergence.

Eigenspectra alone cannot give a complete picture of the accuracy of the reconstruction, so we examine next the spatial POD modes obtained from the

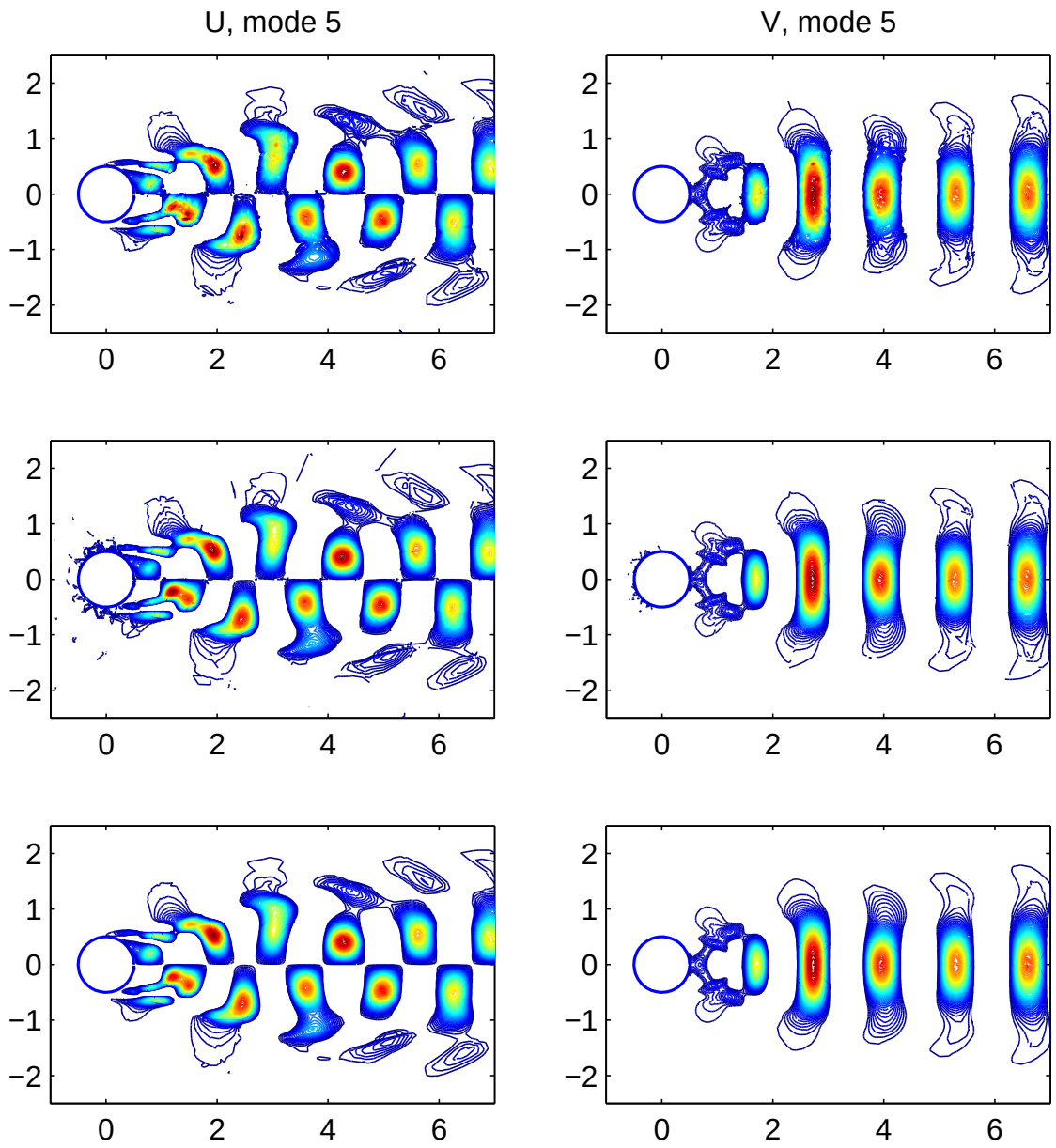

Fig. 5. High temporal resolution: Comparison of the 5^{th} POD mode at $Re = 500$ with 50% gappiness. From top to bottom: POD-1, Kriging interpolation and DNS field. Left: Streamwise velocity. Right: Crossflow velocity.

reconstructed fields. It was demonstrated in [6] that achieving an accurate (temporal) eigenspectrum does not necessarily imply accurate reconstruction of the spatial modes, especially for large gappiness. For the most energetic modes, there is good agreement between all three reconstruction approaches (POD-1, POD-2, Kriging) with the original complete data, as shown in figure 5 for the fifth mode. Here we do not show separately the POD-2 results as they have no visible differences with the POD-1 results. Note that the Kriging results are somewhat noisy compared to the original data unlike the POD results. Next, we plot results for a typical high mode (mode 18th) in figure 6.

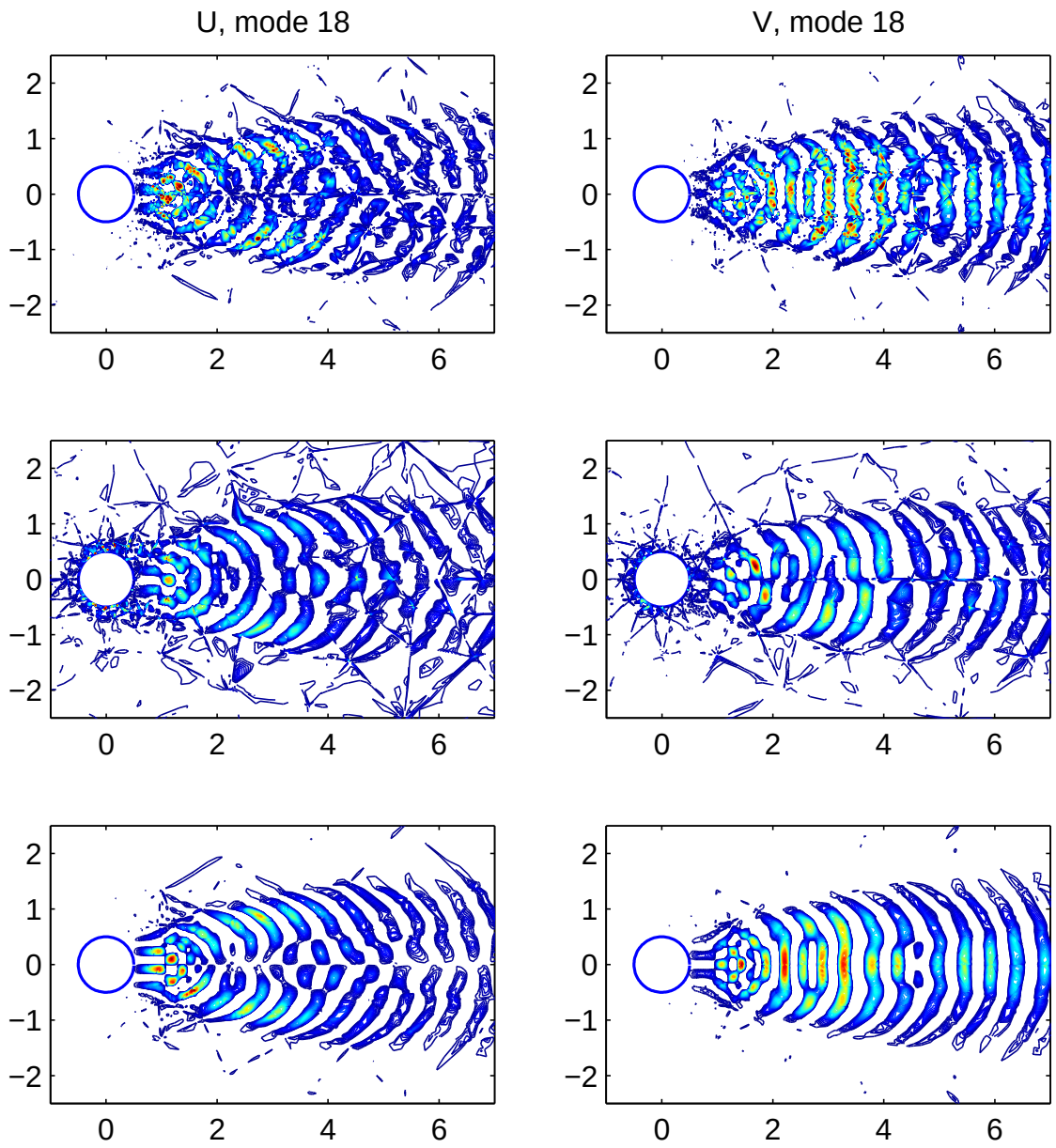

Fig. 6. High temporal resolution: Comparison of the 18^{th} POD mode at $Re = 500$ with 50% gappiness. From top to bottom: POD-1, Kriging interpolation and DNS field. Left: Streamwise velocity. Right: Crossflow velocity.

Unlike the low modes, here Kriging interpolation gives erroneous results, as the actual POD mode has opposite symmetries compared to what is obtained with Kriging interpolation. However, the POD-based algorithms produce correct results compared to the POD modes obtained from the whole DNS field.

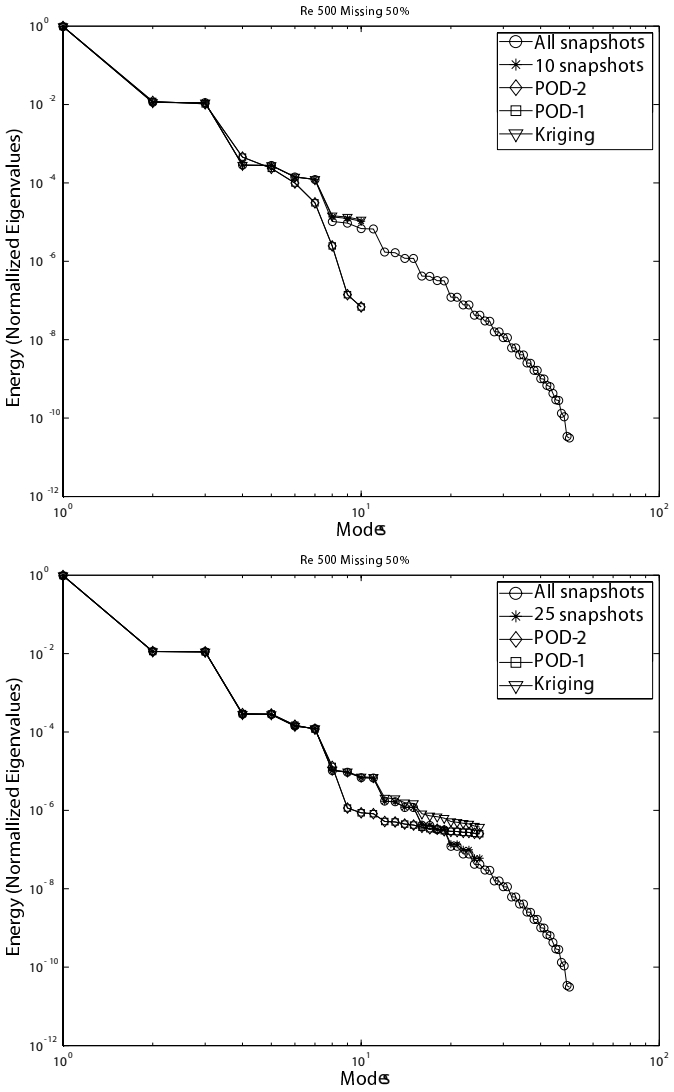

Fig. 7. Low temporal resolution: Comparison of eigenspectra of the recovered data for 50% gappiness at $Re = 500$. Top: $N = 10$ snapshots. Bottom: $N = 25$ snapshots.

4.1.3 Low temporal resolution

Temporal resolution affects greatly the accuracy of the POD based reconstruction procedures. We study here the case in which we assume that we do not have sufficient temporal resolution. This case is studied for the same gappy data sets but now we drop some of the snapshots so that there are only $N = 10$ or $N = 20$ equally spaced snapshots per vortex shedding period for $Re = 100$ and $N = 10$ or $N = 25$ snapshots per vortex shedding period for $Re = 500$. Here we present results for the higher Reynolds number in figure 7. Unlike the high temporal resolution case, both POD-1 and POD-2 perform poorly compared to the Kriging interpolation. We note that in the case of $Re = 500$ with 10 snapshots, the eigenspectrum of the two POD versions matches the actual spectrum only up to $M_0 = 3$ modes whereas that obtained using the

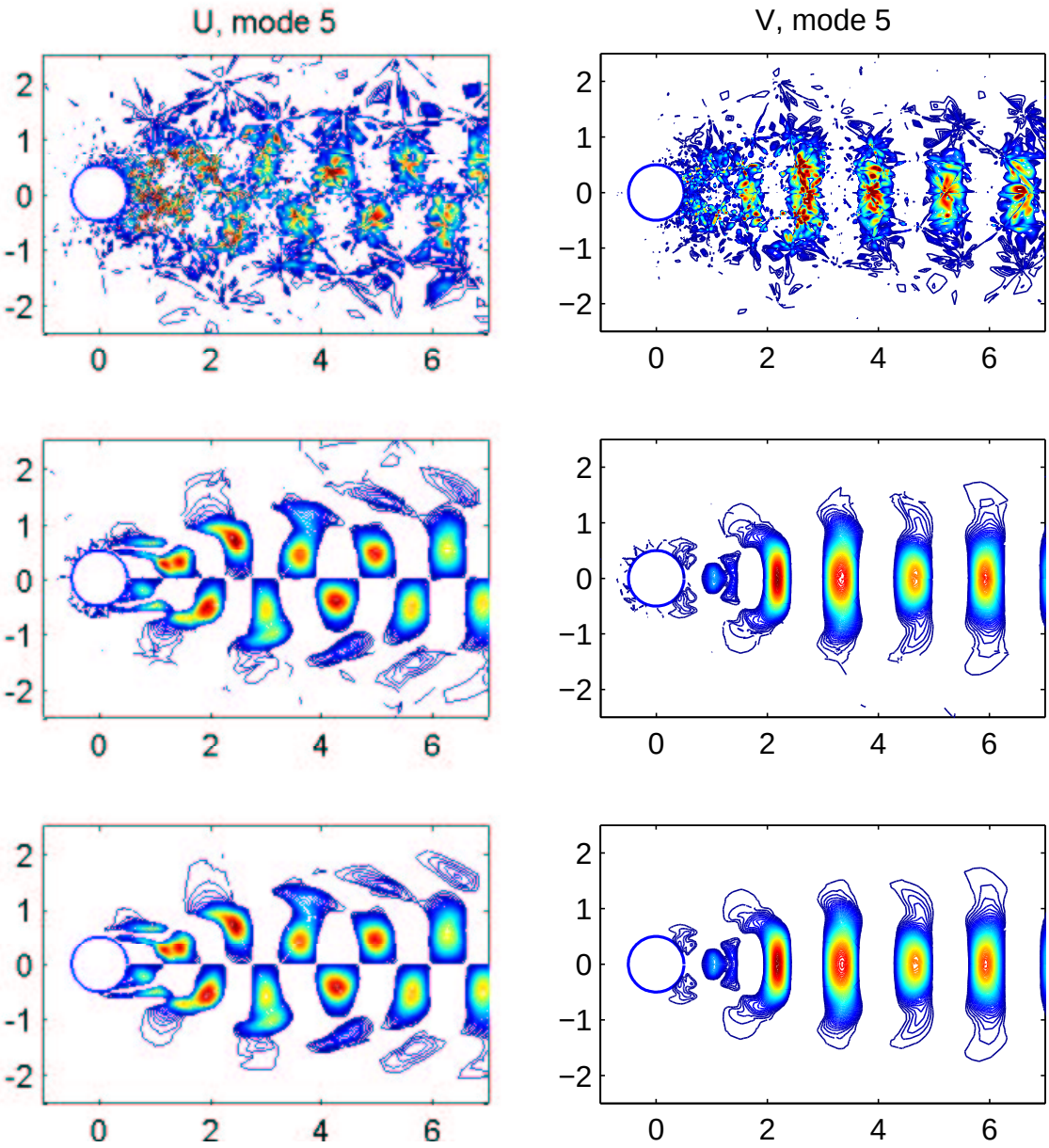

Fig. 8. Low temporal resolution: comparison of the fifth POD mode of $Re = 500$ with 10 snapshots for 50% gappiness. From top to bottom: POD-1, Kriging interpolation, DNS field. Left: Streamwise velocity. Right: Crossflow velocity.

Kriging method is accurate up to $M_0 = 7$ modes. We have also included in the plot the eigenspectra of the complete fields corresponding to the entire set of snapshots in order to indicate that there is almost a total overlap of the corresponding eigenspectra. Finally, in figure 8 we plot contours of the fifth spatial mode confirming that indeed Kriging is more effective for flow data corresponding to low temporal resolution. We do not present any higher modes here as they are very noisy.

4.1.4 *Kriging for black zones*

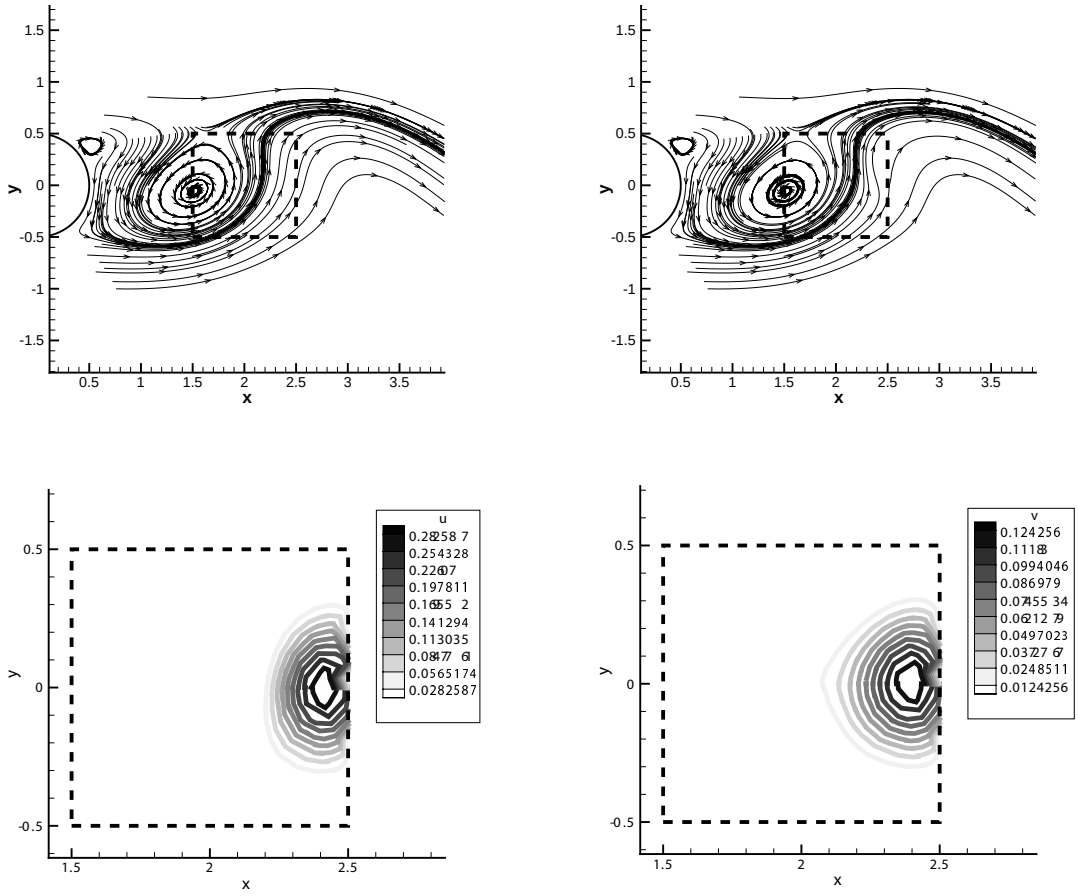

Fig. 9. Top: Kriging results for a black zone $1.5D$ downstream from the cylinder. Left: Original field. Right: Kriging reconstruction. Bottom: Error due to Kriging. Left: Error in streamwise velocity. Right: error in crossflow velocity. (The freestream velocity is $U_\infty = 1$).

The POD-based reconstruction procedures cannot recover any data for the black zones, i.e. regions in the domain for which we have missing data at *all* times. However, Kriging interpolation can fill in these black zones, so here we evaluate its effectiveness. To this end, we will employ again the 2D simulation data in flow past a cylinder. Results with available *experimental* data based on particle-image-velocimetry (PIV) were obtained in [12] for a more complex flow and similar conclusions as the ones presented next were drawn.

First, we consider uniform flow past a circular cylinder, as above, and we remove from the domain some regions. The first removed region is a square area of width D. It is located at $1.5D$ downstream from the cylinder. The second removed region is a square area with $2D$ width located at $10.5D$ downstream of the cylinder. We have removed the same area for all 40 snapshots and 50 snap-

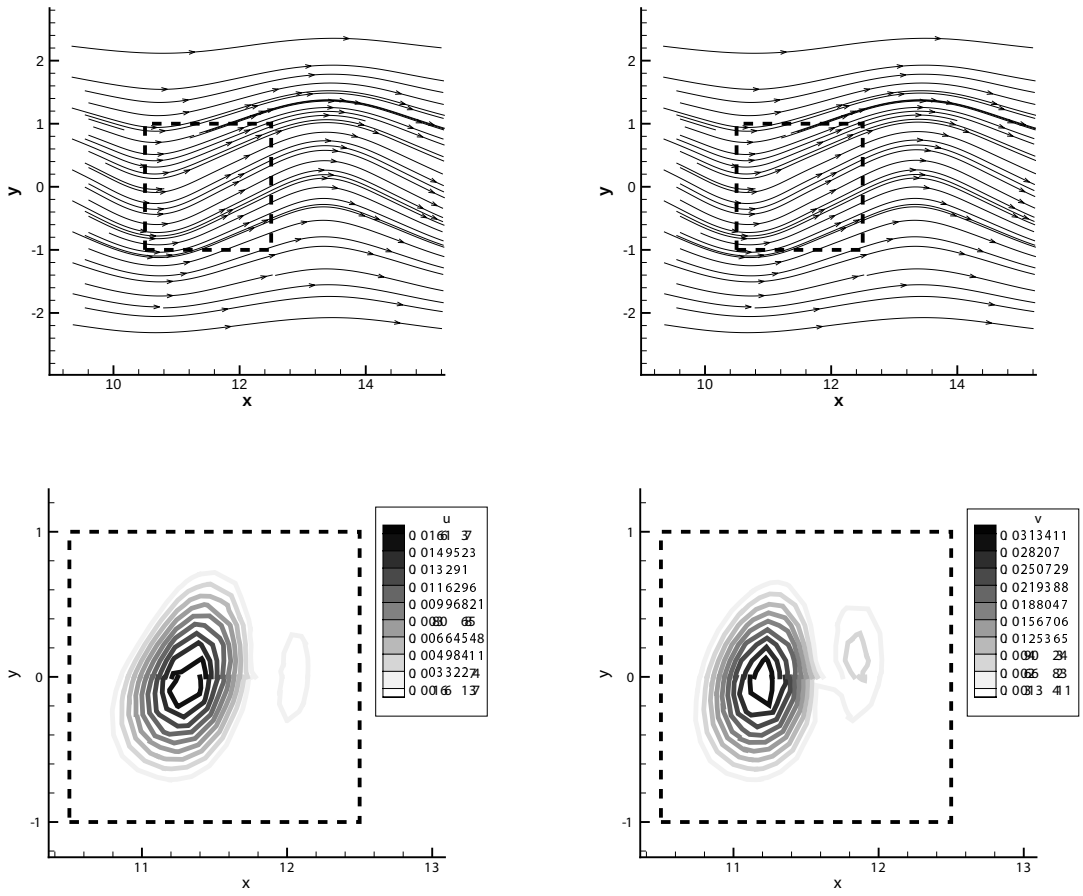

Fig. 10. Top: Kriging results for a black zone $10.5D$ downstream from the cylinder. Left: Original field. Right: Kriging reconstruction. Bottom: Error due to Kriging. Left: Error in streamwise velocity. Right: error in crossflow velocity. (The freestream velocity is $U_\infty = 1$).

shots for $Re = 100$ and $Re = 500$. The Kriging interpolation now is performed separately on two bigger square domains containing the aforementioned black zones with width of $2D$ and $4D$.

The results at $Re = 500$ lead to same conclusion as the results from $Re = 100$, so here we present results from $Re = 100$. We plot the streamlines of the first snapshot of the data from $Re = 100$ in figure 9. Shown here is the case where the black zone is at $1.5D$ downstream (the dashed box represents the black zone.) We see that the results obtained from the Kriging interpolation seem to be in reasonable agreement with the real field. We can quantify the error by plotting in figure 11 (left) the streamwise velocity profile at $x = 2$, which is a location right through the middle of the black zone; the error is noticeable but small.

Next, we examine the case for which the black zone is at $10.5D$ downstream

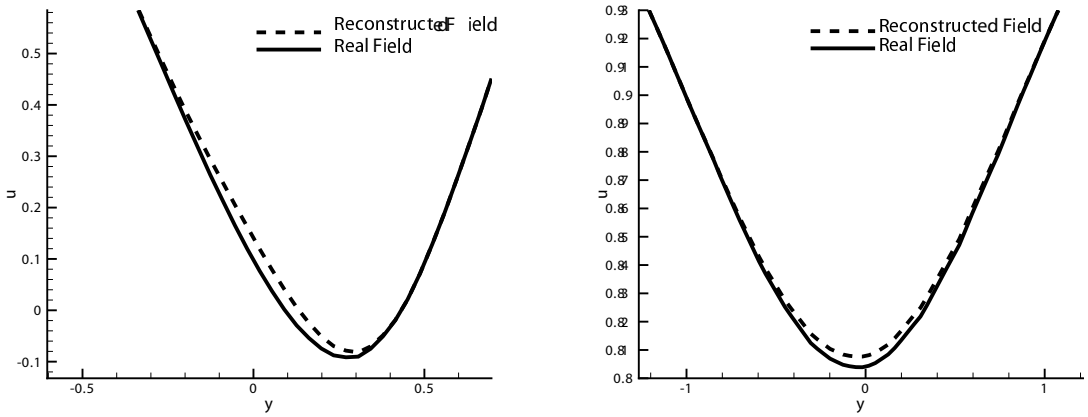

Fig. 11. Streamwise velocity profile at $x = 2$ (left) and $x = 12$ (right) for $Re = 100$ for black zones located at 1.5D and 10.5D, respectively, downstream from the cylinder.

from the cylinder, see figure 10. The results from this figure in conjunction with the velocity profile plotted in figure 11 (right) demonstrate that the prediction of the Kriging interpolation is more accurate in this case. This should be expected as the first black zone is in the region of *absolute instability*, which effectively defines the von Karman street behind the cylinder, see [19].

4.2 Three-Dimensional Gappy Fields

In the first subsection we continue the comparison between POD-based reconstruction and Kriging interpolation for three-dimensional flow. In the second subsection we investigate how the accuracy of Kriging interpolation is affected by the correlation kernel we select; we do this for the turbulent wake. Due to excessive memory requirements in the POD-based reconstruction we could not process the data with our serial reconstruction code for the turbulent field, so a comparison in this case was not possible. Kriging, on the other hand, works on one field at at time.

4.2.1 Periodic limit cycle

At Reynolds number $Re = 185$ the flow past a cylinder bifurcates to a stable three-dimensional time-periodic state [20]. We want to investigate the performance of the aforementioned reconstruction techniques for gappy data sets of this three-dimensional flow. We have shown in the previous section that at least for the cylinder flow the two POD-based reconstruction procedures perform similarly, so here we present comparisons of POD-1 against Kriging interpolation.

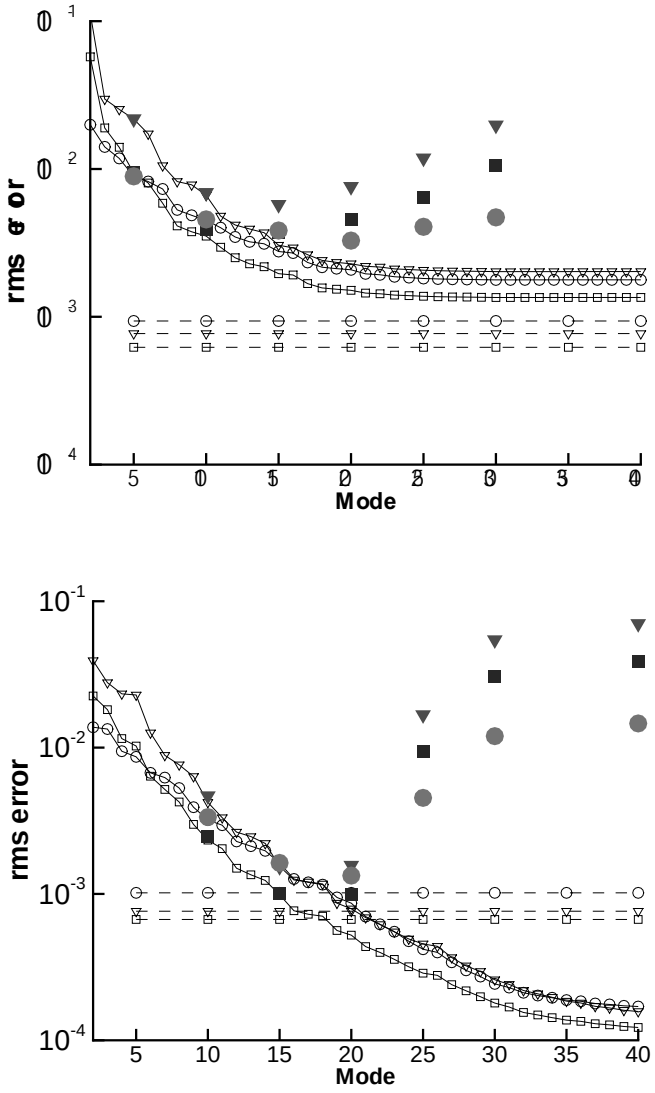

Fig. 12. $Re = 185$ and 25% gappiness: Root-mean-square error of the reconstructed velocity components as a function of the number of modes employed in the reconstruction. Solid line: POD-1, dashed line: Kriging, filled symbols: original Everson-Sirovich procedure. Top: $N = 40$, Bottom: $N = 80$. Square: streamwise component; Triangle: crossflow component; Circle: spanwise component.

In order to investigate the accuracy of each reconstruction procedure in detail we employ the standard root-mean-square error (*rms*) for each snapshot as follows:

$$rms^M(u) = \sqrt{\frac{1}{N_T}\sum_{i=1}^{N_T}\left[u_c^M(x,y,z) - u(x,y,z)\right]^2} \tag{10}$$

where N_T is the total number of grid points in the flow field, u_c^M denotes the

reconstructed streamwise velocity component via M modes, while $u(x, y, z)$ denotes the "true" velocity component obtained from the DNS before the random data-discarding process. The *rms* error for the crossflow and spanwise velocity components are defined similarly.

Figure 12 shows the root-mean-square error of the reconstructed velocity components as a function of mode number employed in the POD-1 method and the Kriging method. The gappy field considered corresponds to 25% gappiness percentage. The POD-1 method exhibits monotonic convergence as the number of modes is increased. The Kriging procedure is independent of the number of snapshots since it is employed for each snapshot individually. We have also included (with symbols) the original (non-iterative) POD-1 procedure of Everson & Sirovich, which converges initially but diverges above a certain number of modes. This is a typical behavior we have observed for the two-dimensional flow as well, see also [6]. We also observe here that the relative performance of POD-1 against Kriging depends strongly on the temporal resolution, i.e., the number of available snapshots. Specifically, by increasing the number of snapshots from $N = 40$ to $N = 80$, the error for POD-1 drops more than one order of magnitude. We also note that while there is almost no difference in the (overlapping) POD spectrum of full (non-gappy) flow field between 40 and 80 snapshots, it is crucial to employ a sufficient number of snapshots for an accurate reconstruction of the gappy data.

We now examine how the performance of the two reconstruction methods changes as the gappiness increases. Figure 13 shows the root-mean-square error of the reconstructed velocity components as a function of mode number for 50% and 75% gappiness. Comparing these results with the results of figure 12 suggests that for increased gappiness levels (up to 75%, possibly more), Kriging outperforms the POD-based approach. The convergence of POD-1 as a function of the number of modes employed in the reconstruction is rather complex: While for the low gappiness percentage (figure 12) , all modes, including higher modes, contribute to reducing the error, for larger gappiness (50% and 75%), higher order modes do not actually contribute to reducing the error. This can be attributed to the unresolved higher (spatial) modes due to the insufficient number of snapshots for larger gappiness percentage. Since each snapshot provides some extra information of the dynamics of the flow, by increasing the number of snapshots beyond $N = 80$ for 75% gappiness, it may be possible to further improve the accuracy of the POD-1 method.

Figure 14 shows a comparison of eigenspectra obtained using the POD-1 method and the Kriging procedure. Note that the eigenspectrum obtained by the Kriging procedure does not depend on the number of modes while for the POD-1 method the eigenspectrum (like the *rms* error) is a function of the number of modes, and can be improved significantly by increasing the number of modes for reconstruction. Figure 14 shows that for all three gappiness

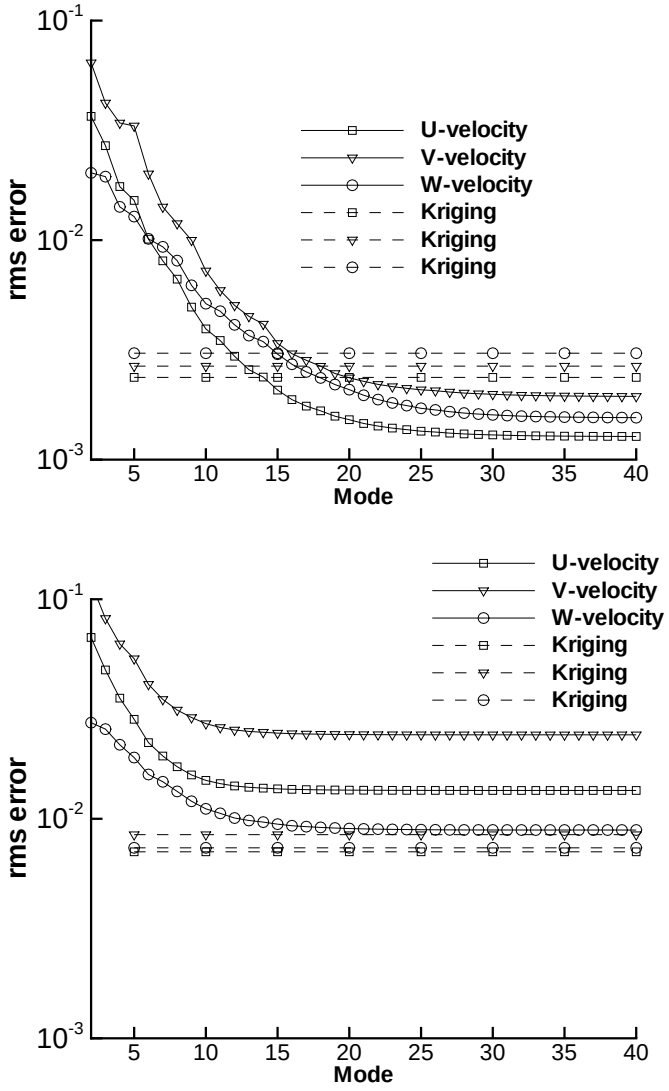

Fig. 13. Root-mean-square error of the reconstructed velocity components as a function of mode number ($N = 80$). For legend see caption of figure 12. Top: 50% gappiness. Bottom: 75% gappiness.

percentages considered, more *temporal modes* are resolved using the POD-1 procedure. Specifically, there are apparent large deviations for the Kriging method for higher modes as seen in figure 14. In other words, it appears that the POD-1 procedure follows the "true" spectrum more closely. However, this is true for the low gappiness percentage only. Upon closer inspection of the results for the 75% gappiness, we see in figure 15 that there are considerable deviations in the *most energetic* temporal modes obtained using the POD-1 method whereas Kriging follows the most important part of the "true" spectrum, i.e. the low more energetic modes, more accurately.

Table 1 shows the relative error of the first three temporal modes for differ-

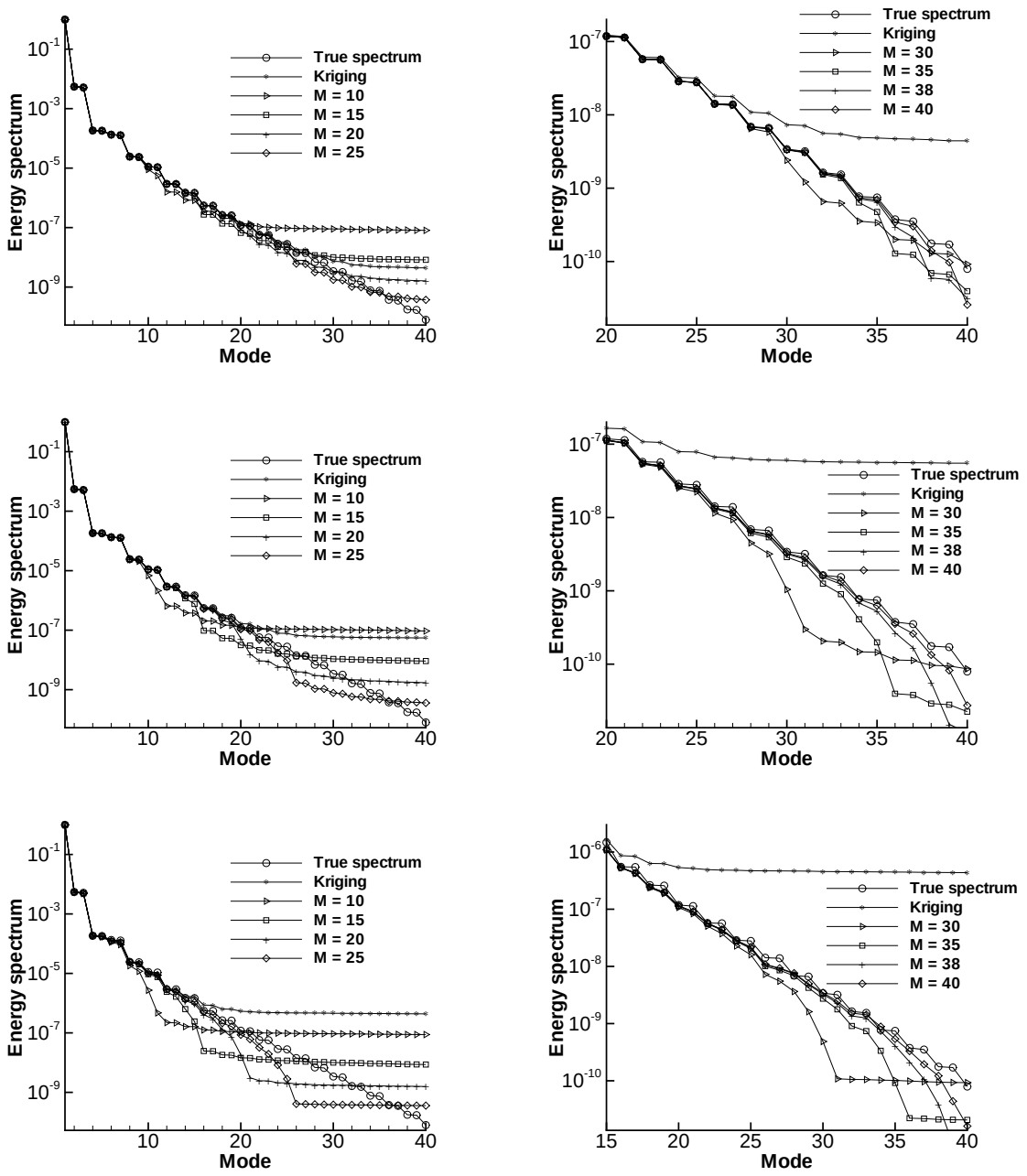

Fig. 14. Comparison of eigenspectra for different number of modes. On the right a close-up view with more number of modes is presented. From top to bottom 25% gappiness, 50% gappiness and 75% gappiness.

ent gappiness percentages. At 75% gappiness percentage, the Kriging method captures the first three temporal modes better than the POD-1 procedure by about an order of magnitude whereas for lower gappiness the POD-1 method is more accurate. Overall, the errors in eigenspectra are consistent with the *rms* errors we presented in earlier plots.

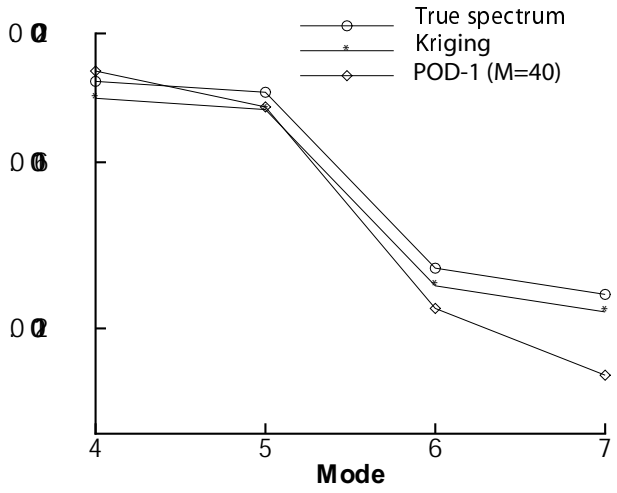

Fig. 15. Close-up of eigenspectra (shown in figure 14) for the case of 75% gappiness.

M	25%		50%		75%	
	POD-1 procedure	Kriging	POD-1 procedure	Kriging	POD-1 procedure	Kriging
1	0	$1.0\text{x}10^{-5}$	$2.2\text{x}10^{-4}$	$2.7\text{x}10^{-4}$	$3.69\text{x}10^{-2}$	$3.74\text{x}10^{-3}$
2	0	0.0	0	$2.5\text{x}10^{-2}$	2.3	0.4
3	0	$1.0\text{x}10^{-5}$	$1.4\text{x}10^{-2}$	$3.7\text{x}10^{-2}$	4	0.4

Table 1
Relative error of the first three temporal modes (eigenvalues) compared to the "true" eigenvalues.

We now turn our attention to the quality of the reconstructed velocity fields. For the smallest gappiness we considered, i.e. 25%, all POD modes up to the twenty-first mode are resolved accurately by both methods. For higher order modes, e.g., the thirty-first mode, there is a large difference between the two methods with POD-1 much more accurate than Kriging. Despite the inaccuracy in the higher modes, the reconstructed flow fields are visually identical.

We now examine the reconstructed fields at gappiness 50% for which the POD-1 eigenspectrum is slightly more accurate than that of Kriging interpolation. Contours of all three velocity components are plotted in figure 16 at a fixed z-plane perpendicular to the cylinder axis. In addition to the reconstructed fields, we also present the 50% gappy data and the (non-gappy) DNS data, for comparison. The left column represents the streamwise component, the middle column represents the cross flow component and the right column represents the spanwise component of the flow. The POD-based approach results in a slightly better reconstruction than the Kriging procedure. For example, the spanwise velocity component error of the Kriging procedure is twice as big

as the POD-1 procedure (0.003054 versus 0.001553) and this leads to visibly "noisy" contours of spanwise velocity component depicted in figure 16 obtained using the Kriging procedure. Figures 17 and 18 show spanwise variation of the reconstructed crossflow and spanwise velocity fields, respectively. In general, both methods perform very well in capturing spanwise distributions, even though some slight deviations from the "true" field are noticeable for the Kriging procedure under closer examination, consistent with the respective error values.

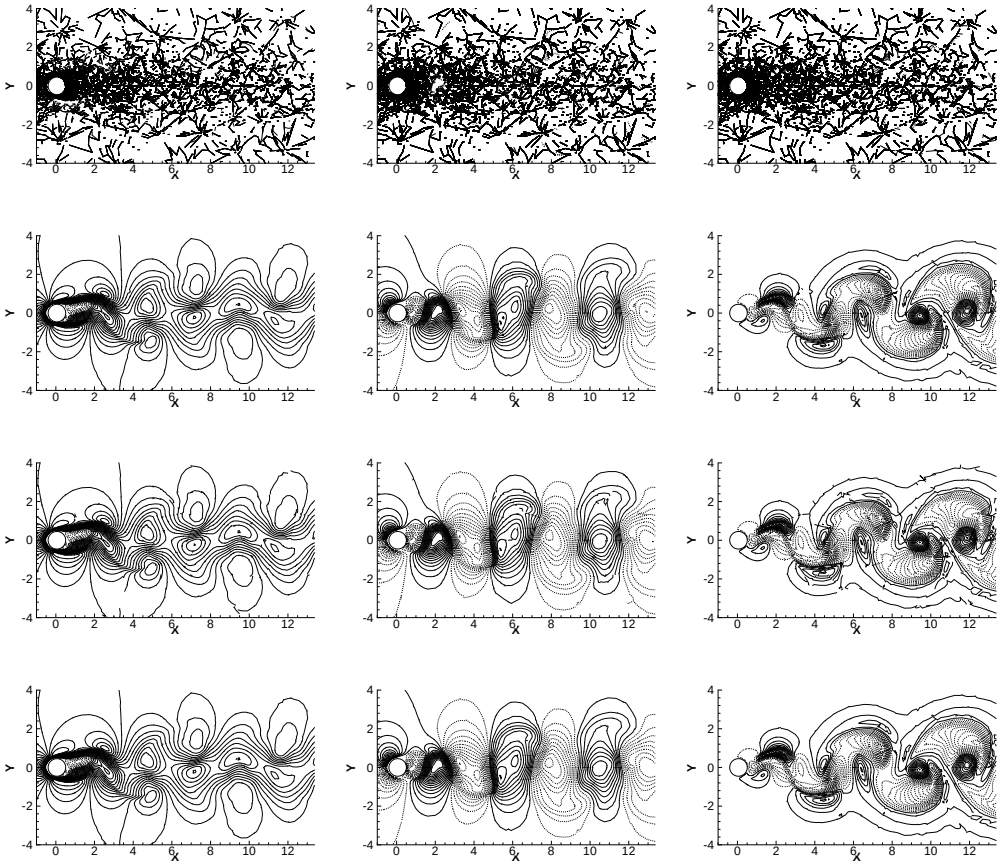

Fig. 16. Reconstruction of 50% gappy data via the POD-1 and Kriging methods. From top to bottom, gappy data field, POD-1, Kriging, and original DNS data. The instantaneous ninth snapshot at $z = 1$ plane is shown. Left: streamwise; Middle: crossflow, and Right: spanwise components.

Finally, we present the reconstructed fields for very large gappiness, namely 75%. Unlike the 25% and 50% gappy data we presented previously, for the 75% gappy data set, none of the modes obtained by the POD-1 are resolved accurately; they all contain noise, which increases drastically with increasing mode number. On the other hand, the lower most energetic modes (e.g., up to tenth mode) extracted by Kriging are resolved accurately. The effect of unresolved modes (both temporal and spatial) is evident on the reconstructed flow field by the POD-1, as shown in figure 19, where the crossflow velocity

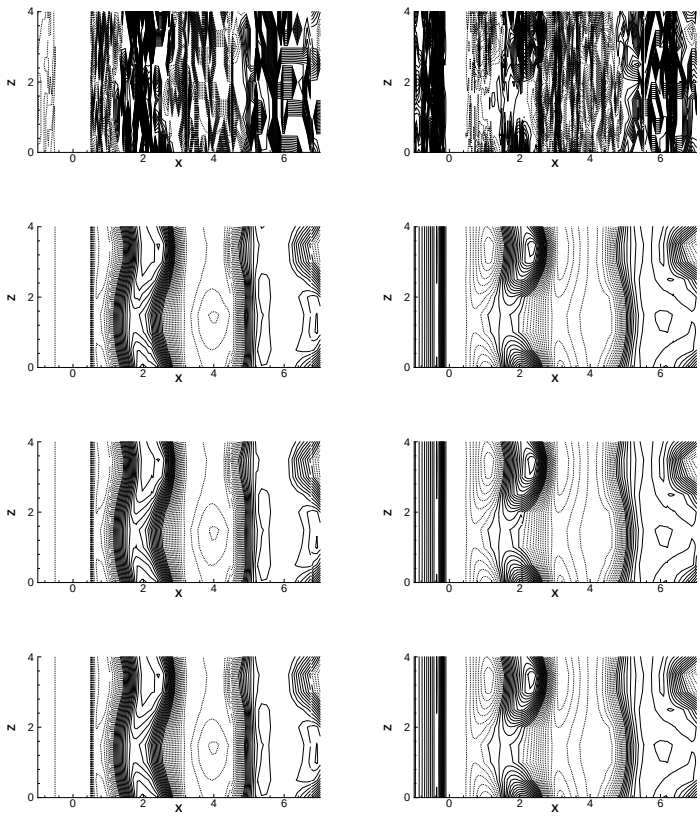

Fig. 17. Reconstruction of the crossflow velocity component for 50% gappy data. From top to bottom, gappy data field, POD-1, Kriging, and original DNS data. Left column: y= 0 plane, right column: $y = 0.5$ plane. (ninth snapshot)

component is plotted.

4.3 Kriging: Choice of Correlation Function

We have demonstrated so far that the Kriging procedure is quite effective for laminar flows for which the velocity fields are very smooth. In this section we consider turbulent flow fields obtained in simulations described in [18] for flow past a cylinder at Reynolds number $Re = 10,000$. A mesh consisting of 9272 triangular elements was employed with Jacobi polynomial basis of of fifth order, while 64 Fourier planes were employed along the span. The Kriging procedure is applied in each element in order to recover the missing data. Here, we compare two types of kernels corresponding to different smoothness at the origin:

(1) The Gaussian kernel, which is the kernel we have used in all cases so far at lower Reynolds number, and
(2) The exponential kernel which is non-smooth at the origin.

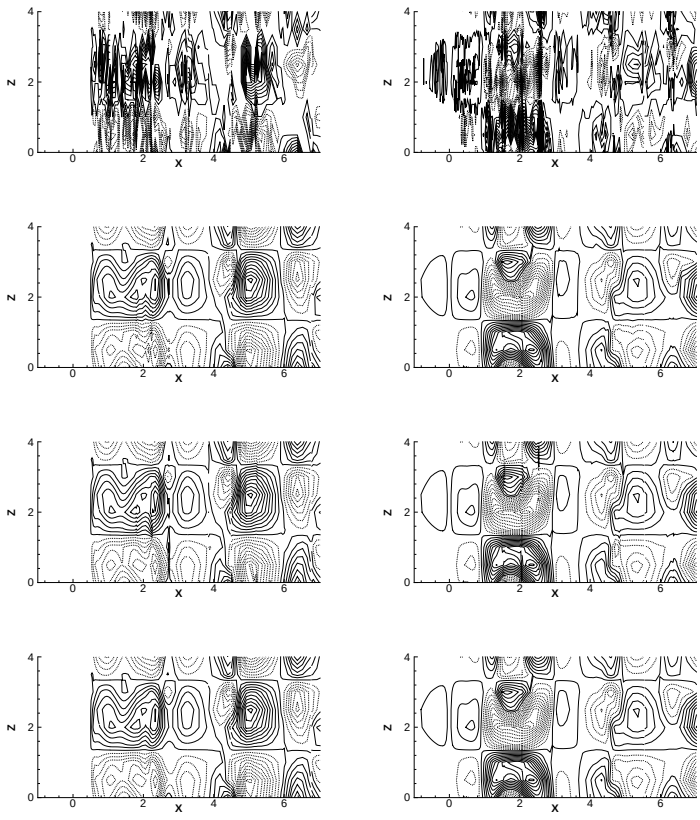

Fig. 18. Reconstruction of the spanwise velocity component for 50% gappy data. From top to bottom, gappy data field, POD-1, Kriging, and original DNS data. Left column: $y = 0$ plane, right column: $y = 0.5$ plane. (ninth snapshot)

A summary of our findings is presented in table 2. Reconstruction of the streamwise velocity exhibits the largest errors. For low gappiness (25%), the Gaussian kernel is clearly more effective than the exponential kernel. However, for the higher gappiness cases, the two kernels are equally effective as the errors in the reconstruction are about the same.

5 Summary

We have presented three different algorithms in reconstructing gappy velocity fields for unsteady flows. The first two algorithms are based on proper orthogonal decomposition (POD) and require the entire set of snapshots. They are different in the way the unknown temporal modes are computed: Algorithm POD-1 involves the solution of a least-squares system whereas algorithm POD-2 is less rigorous. The third method is based on Kriging interpolation, it is local, and it is applied to each snapshot separately. In the current study we have created artificially gappy data sets obtained from direct numerical simulations of flow past a circular cylinder in two- and three-dimensions. We have

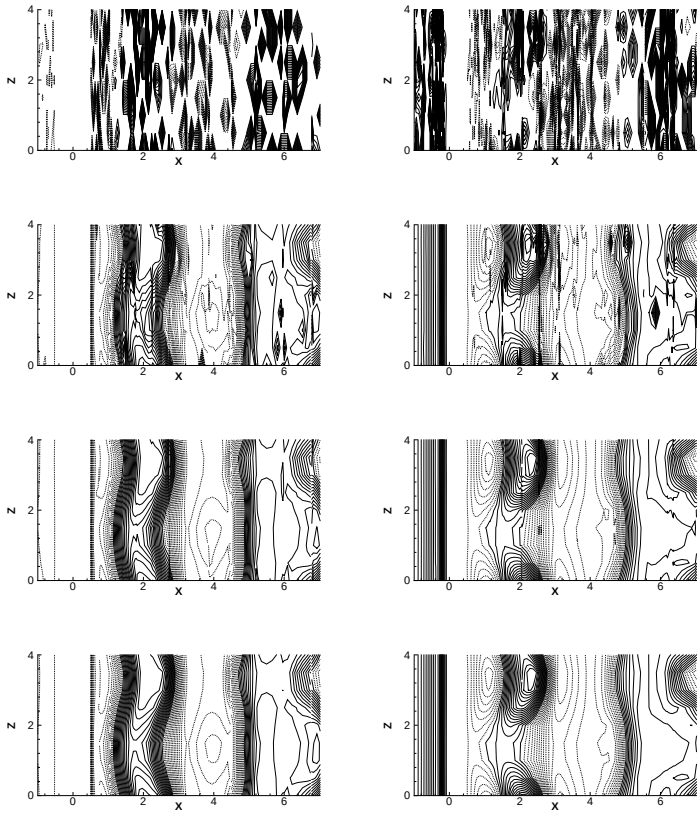

Fig. 19. Reconstruction of the cross flow velocity component for 75% gappy data. From top to bottom, gappy data field, POD-1, Kriging and original DNS data. Left column: $y = 0$ plane, right column: $y = 0.5$ plane. (ninth snapshot)

also worked with experimental data obtained using particle image velocimetry (PIV) and similar results, as in the current study, were obtained, see [12]. We summarize here the main findings from our study:

- Kriging interpolation is an effective way of recovering missing data in unsteady flows even in sensitive regions, e.g. regions of absolute instability.
- For high temporal resolution (i.e., many snapshots), POD-based reconstruction is more accurate than Kriging interpolation; however, for low temporal resolution Kriging is more effective.
- For small gappiness in the flow field, POD-based reconstruction is more accurate than Kriging; however, for large gappiness Kriging is more accurate.
- The two versions of POD-based reconstruction exhibit the same accuracy for modest levels of gappiness, but POD-2 is much more efficient than POD-1. However, for large gappiness and noisy gappy data POD-1 is more accurate.
- In Kriging interpolation, the Gaussian correlation kernel is more effective than the exponential kernel but for large gappiness similar results are obtained.

Kernels	Rms(u)		
	25%	50%	75%
Gaussian type	0.00688	0.01655	0.03043
Exponential type	0.00924	0.01799	0.03120
Kernels	Rms(v)		
	25%	50%	75%
Gaussian type	0.00672	0.01614	0.02964
Exponential type	0.00905	0.01756	0.03042
Kernels	Rms(w)		
	25%	50%	75%
Gaussian type	0.00566	0.01458	0.02886
Exponential type	0.00801	0.01637	0.02992

Table 2
Error (rms) in velocity components using two different Kriging kernels for a turbulent field.

The flow we have examined exhibits low-dimensionality and therefore the reconstruction methods are effective even for large gappiness, e.g. 75%. We have also worked with relatively smooth data, whereas, in general, noise may be present in the gappy data sets. In future work, it will be of interest to apply the techniques presented here to other flows, e.g. to wall-bounded turbulence, in order to gain more experience with data recovery techniques for such fluid mechanics problems.

Acknowledgments

H. Gunes gratefully acknowledges the financial support of The Scientific and Technical Research Council of Turkey (TUBITAK). S. Sirisup gratefully acknowledges the DPST (Development and Promotion of Science and Technology Talents) project from Thailand for providing his scholarship during his graduate studies at Brown University. We would like to thank Dr. Steve Dong for helpful discussions. This work was supported by NSF (Dr. F. Darema) and ONR (Dr. T.F. Swean), and computations were performed at the facilities of NCSA (University of Illinois at Urbana-Champaign) and at DoD's NAVO MSRC.

References

[1] R. Little, D. Rubin, Statistical Analysis with Missing Data, Wiley Interscience, 2002.

[2] F. Yates, The analysis of replicated experiments when the field results are incomplete, Emp. J. Exp. Agri. 1 (1933) 129.

[3] M. Stein, Interpolation of Spatial Data: Some Theory for Kriging, Springer-Verlag, 1999.

[4] R. M. Everson, L. Sirovich, The Karhunen-Loève transform for incomplete data, J. Opt. Soc. Am., A 12(8) (1995) 1657–1664.

[5] B.-T. Tan, K. Willcox, M. Damodaran, Applications of proper orthogonal decomposition for inviscid transonic aerodynamics, AIAA paper 2003-4213, 21st Applied Aerodynamics Conference, 23-26 June.

[6] D. Venturi, G. Karniadakis, Gappy data and reconstruction procedures for flow past a cylinder, J. Fluid Mech. 519 (2004) 315–336.

[7] J. M. Beckers, M. Rixen, EOF calculations and data filling from incomplete oceanographic datasets, J. Atmos Ocean Tech 20(12) (2003) 1839–1856.

[8] X. Ma, G. Karniadakis, H. Park, M. Gharib, DPIV-driven simulation: A new computational paradigm, Proc. R. Soc. Lond. A 459 (2003) 547–565.

[9] X. Ma, G. Karamanos, G. Karniadakis, Dynamics and low-dimensionality of the turbulent near-wake, J. Fluid Mech. 410 (2000) 29–65.

[10] N.-Z. Cao, N. Aubry, Numerical simulation of wake flow via a reduced system, Proc. ASME Fluids Engineering Conference, Washington D.C.

[11] A. Deane, I. Kevrekidis, G. Karniadakis, S. Orszag, Low-dimensional models for complex geometry flows: Application to grooved channels and circular cylinders, Phys. Fluids A 3 (10) (1991) 2337–2354.

[12] S. Sirisup, Issues in low-dimensional modeling of unsteady flows: convergence, asymptotic stability and reconstruction procedures, Ph.D. thesis, Division of Applied Mathematics, Brown University (2005).

[13] J. Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, Design and analysis of computer experiments, Stat Sci 4(4) (1989) 409–423.

[14] S. Lophaven, H. Nielsen, J. Søndergaard, DACE - A Matlab Kriging Toolbox, Version 2.0., Report IMM-REP-2002-12, Informatics and Mathematical Modelling, DTU, 2002.

[15] S. Lophaven, H. Nielsen, J. Søndergaard, Aspects of the Matlab Toolbox DACE, Report IMM-REP-2002-13, Informatics and Mathematical Modelling, DTU, 2002.

[16] G. Karniadakis, S. Sherwin, Spectral/hp Element Methods for CFD, Oxford University Press, 1999.

[17] X. Ma, G. Karniadakis, A low-dimensional model for simulating 3d cylinder flow, J. Fluid Mech. 458 (2002) 181–190.

[18] S. Dong, G. Karniadakis, DNS of flow past a stationary and oscillating cylinder at Re=10,000, J. Fluid & Struct. in press.

[19] G. Karniadakis, G. Triantafyllou, Frequency selection and asymptotic states of laminar wakes, J. Fluid Mech. 199 (1989) 441.

[20] C. Williamson, Vortex dynamics in the cylinder wake, Ann. Rev. Fluid Mech. 28 (1996) 477–539.